

The process of most recent common ancestors in an evolving coalescent

BY P. PFAFFELHUBER* AND A. WAKOLBINGER

*Ludwig-Maximilian University Munich
and Goethe-University Frankfurt*

Abstract

Consider a haploid population which has evolved through an exchangeable reproduction dynamics, and in which all individuals alive at time t have a most recent common ancestor (MRCA) who lived at time A_t , say. As time goes on, not only the population but also its genealogy evolves: some families will get lost from the population and eventually a new MRCA will be established. For a time-stationary situation and in the limit of infinite population size N with time measured in N generations, i.e. in the scaling of population genetics which leads to Fisher-Wright diffusions and Kingman's coalescent, we study the process $\mathcal{A} = (A_t)$ whose jumps form the point process of time pairs (E, B) when new MRCAs are established and when they lived. By representing these pairs as the entrance and exit time of particles whose trajectories are embedded in the look-down graph of Donnelly and Kurtz (1999) we can show by exchangeability arguments that the times E as well as the times B from a Poisson process. Furthermore, the particle representation helps to compute various features of the MRCA process, such as the distribution of the coalescent at the instant when a new MRCA is established, and the distribution of the number of MRCAs to come that live in today's past.

1 Introduction

The genealogy back to the most recent common ancestor (MRCA) of those currently alive, and especially the time back to the MRCA, has been an ongoing object of interest in mathematical population genetics, see [Lit75], [Gri80] for early references and [Wak05] for a recent monograph. The limit of effective population size $N \rightarrow \infty$, with time measured in units of N generations, is the scaling in which Kingman's coalescent appears ([Kin82]): in the rescaled time measured backward from a fixed time t , the number of *ancestral lineages* enters from infinity and jumps from k to $k-1$ at rate $\binom{k}{2}$. (Here and below we assume that the population size remains constant in time.) The depth D_t of the *coalescent tree*, that is the rescaled time it takes the number of ancestral lineages to decrease from ∞ to 1, is then a sum of exponentially distributed random variables with mean $\binom{k}{2}^{-1}$, $k = 2, 3, \dots$, and consequently has expectation 2.

With the population evolving further, also its genealogical relationships given by the coalescent tree change. In this study we are interested in the time evolution of one particular characteristics of the genealogy, that is, the time $A_t = t - D_t$ when the MRCA of the population at time t lived. We will refer to $\mathcal{A} = (A_t)_{t \in \mathbb{R}}$ as the *MRCA process*.

At any time t the total population consists of two oldest families, which stem from the two oldest lines of descent dating back to the MRCA who lived at time A_t . These two families will coexist for a while after time point t , and during this time interval the path of the MRCA process \mathcal{A} stays constant. At some random time $E_t > t$, one of the two families will go extinct and the other one will fixate in the population. The MRCA of this surviving family must be more recent

*Travel support from DFG, Bilateral Research Group FOR 498.

AMS 2000 subject classification. 60K35 (Primary) 92D25 (Secondary).

Keywords and phrases. Kingman's Coalescent, Look-down process, most recent common ancestor

than A_t , which amounts to a jump of the MRCA process at time E_t . Consequently, at time E_t , the next MRCA is established, and the time when this next MRCA lives is $B_t := A_{E_t}$. In other words, the path of the process \mathcal{A} is constant as long as the two currently oldest families coexist in the population, and jumps from A_t to B_t at time E_t when one of the two families fixates.

The MRCA process is embedded in the genealogy of the population which is assumed to evolve in a time stationary way. For a finite population consisting of N individuals, a way to construct the genealogy comes with the graphical representation of the Moran model: for each ordered pair (i, j) of indices $i \neq j \in \{1, \dots, N\}$, an exponential clock rings at rate $1/2$, and whenever this happens, the individual with index j dies and is replaced by an offspring of the individual with index i . This results in a partitioning of $\mathbb{R} \times \{1, \dots, N\}$ into coalescing ancestral lineages, from which one can read off a version of the MRCA process for N individuals. This process clearly inherits time stationarity from the evolution of the population.

The Moran dynamics is exchangeable with respect to the individuals' indices. In contrast, the look-down process introduced by Donnelly and Kurtz (1999), which is the basic tool in our study and will be reviewed in Section 2, arranges the individuals' indices (henceforth referred to as *levels*) at any time according to the persistence of the individuals' offspring in the population: the offspring of an individual at level i outlives the offspring of any contemporary individual at some higher level. For a finite population number N this is achieved as follows: Each level j "looks down" to each smaller level i at rate 1. Whenever this happens, all individuals at levels $j, \dots, N-1$ are pushed one level up, the individual at level N is killed, and the individual at level i spawns a child at level j . The time stationary MRCA process read off from the look-down graph obviously has the same distribution as the time stationary MRCA process read off from the Moran graph.

The look-down process allows a passage to the limit of infinite population size in which the ordering by persistence is preserved. The construction of the *random look-down graph* on $\mathbb{R} \times \mathbb{N}$ proceeds in the very same way as described above, except that there is no killing of individuals at any finite level. Instead, the offspring of an individual at level $i \geq 2$ goes to extinction as soon as this *line of ascent* of the individual is pushed to infinity. All this will be explained in more detail in Section 2.

Because of the ordering by persistence, each MRCA of the population lives at level 1 at some time B at which it gives birth to an individual at level 2. As soon as the offspring of these two individuals fixates in the population, the MRCA is established. Again, because of the ordering by persistence, this happens at the time E when the line of ascent which was pushed at time B from level 2 to level 3 reaches infinity. The process $\mathcal{F} := \{(E, B)\}$, which consists of all pairs of time points when an MRCA is established in the population and when it lived, is a time-stationary point process; we call it the *MRCA point process*. The paths of \mathcal{A} and the point configurations of \mathcal{F} are in an obvious one-to-one correspondence.

The step from time t to the next MRCA, which is established at time E_t and lives at time B_t , and an illustration of the MRCA point process \mathcal{F} are depicted in Figures 1(a) and 1(b) respectively. In both Figures, the left axis contains the times when MRCAs live, and the right axis gives the times when MRCAs are established. The joint distribution of E_t and B_t will be given in Theorem 1 in Section 3. Figure 1(b) displays part of the MRCA point process \mathcal{F} . Remarkably, not only the points B but also the points E form a time stationary Poisson process, see Theorem 2 in Section 4. This will be proved by representing the times $(B, E), (B', E'), \dots$ as the entrance and exit times of particles: the trajectory of a particle is attached to the line of ascent which is pushed from level 2 to level 3 at time B and exits at time E . We will specify the Markovian dynamics of this particle system, compute its equilibrium distribution and show that, whenever a particle exits at some time E , at this very instant the system of (remaining) particles is in equilibrium. This allows to conclude that the waiting time to the next exit time is exponential. The processes \mathcal{A} and \mathcal{F} , however, are not Markov, see Remark 4.1.3.

In Theorem 3 we compute the distribution of the random number $Z_t = \#\{(E, B) \in \mathcal{F} \mid E > t, B < t\}$ of MRCAs that are established after time t and live before time t . In particular, it turns out that the probability that the next MRCA lives in today's future is $\mathbf{P}[Z_t = 0] = 1/3$.

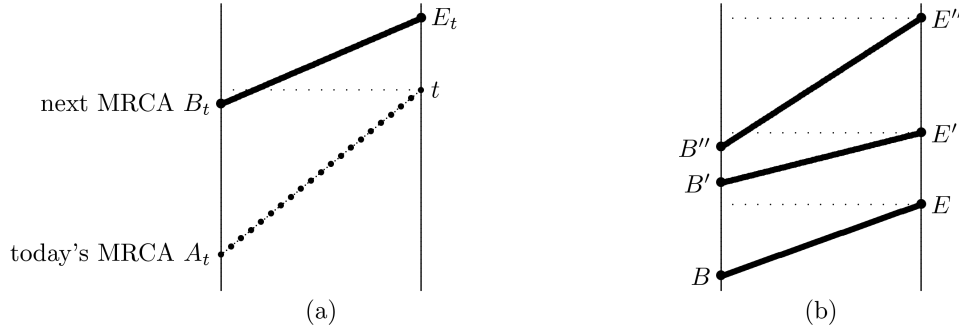


Figure 1: (a) At time A_t the MRCA of the population at time t lived, E_t and B_t are the times when the next MRCA is established and when it lived. (b) MRCAs occur in a time-stationary manner. The dots on the B -axis are time points at which MRCAs lived. The dots on the E -axis are time points at which the MRCA changes.

As noted by [Taj90], the amount of polymorphism in a population is related to the fixation of alleles. When an allele fixates, the MRCA of the population must have changed. At such a fixation time, the full coalescent is unusually short. As neutral mutations fall independently on the branches of the genealogical tree, this means that the amount of polymorphism is low at fixation times.

We start out by reviewing the look-down process in Section 2, describe our results in Sections 3 and 4, point out some relations to population genetics in Section 5 and give the proofs of Theorems 1-3 in Sections 6–8.

2 The MRCA process: a look-down construction

At every time a continuum population which follows a Wright-Fisher (or Fleming-Viot) dynamics has a genealogy given by Kingman's coalescent. The look-down process introduced by Donnelly and Kurtz ([DK99]) not only gives a countable representation of evolving allele frequencies but at the same time stores genealogical relationships of all the individuals alive in the population at all times. Consequently the MRCA process can be read off from the look-down process.

The look-down graph: ancestral lineages, lines of ascent and ordering by persistence

We first give a brief review of the “modified look-down process” ([DK99]); see Figure 2 for a graphical illustration.

Consider the set of vertices

$$\mathcal{V} := \mathbb{R} \times \mathbb{N}.$$

We will refer to the vertex (t, i) as the *individual at time t at level i* . For each ordered pair of levels $i < j$, let \mathcal{P}_{ij} be the support of a (rate one) Poisson point process on \mathbb{R} , all these processes being independent. (In the terminology of Donnelly and Kurtz, at each time $t \in \mathcal{P}_{ij}$, the level j looks down to level i .) Based on the processes \mathcal{P}_{ij} we will construct a random countable partition \mathcal{G} of \mathcal{V} , whose partition elements will be called *lines*. The partition \mathcal{G} will always contain the so-called *immortal line* ι defined by

$$\iota := \mathbb{R} \times \{1\}. \quad (2.1)$$

For each $j > 1$, any point $s_0 \in \bigcup_i \mathcal{P}_{ij}$ initiates a line $G \in \mathcal{G}$ of the form

$$G = ([s_0, s_1) \times \{j\}) \cup ([s_1, s_2) \times \{j+1\}) \cup ([s_2, s_3) \times \{j+2\}) \cup \dots \quad (2.2)$$

with $s_{k+1} > s_k$ for all k . For a line G as in (2.2) with $s_0 \in \mathcal{P}_{ij}$ we say that G is *born at level j by the individual (s_0, i)* . We further say that G is *pushed (one level up)* at times s_1, s_2, \dots and *exits* at time $s_\infty(G) := \lim_{n \rightarrow \infty} s_n$. The times s_k are given for $k = 1, 2, \dots$ by $s_k = \inf\{s > s_{k-1} : s \in \bigcup_{1 \leq \ell < m \leq j+k-1} \mathcal{P}_{\ell m}\}$. Thus, a new line is born at level j at each time t when level j looks down to some level $i < j$. Simultaneously, all the lines having occupied at time $t-$ the levels $j, j+1, \dots$ are pushed one level up. Note that, since the pushing rate increases quadratically in j , the exit time $s_\infty(G)$ is finite a.s.

For each $v \in \mathcal{V}$, we denote by G_v the (unique) element of \mathcal{G} that contains v . The *forward level process* $Y_s^t(i)$, $t \geq s$, initiated by the individual $u = (s, i)$ is given by

$$Y_s^t(i) := \text{level of } G_u \text{ at time } t.$$

The *line of ascent* of individual $u = (s, i)$ is the part of line G_u after time s , that is

$$(t, Y_s^t(i))_{s \leq t < s_\infty(G)}.$$

We say that a line H *descends* from a line G if either $H = G$, or there is a finite sequence of lines $G_1, \dots, G_{n-1} \in \mathcal{G}$ such that G_k is born by an individual in G_{k-1} , $k = 1, \dots, n$, where $G_0 := G$ and $G_n := H$.

The *backward level process* $X_s^t(j)$, $s \leq t$, of an individual $v = (t, j)$ arises by tracing back the level of G_v to the birth time of G_v , then jumping to the level of the individual u from which G_v was born and tracing back the level of G_u to the birth time of (s, i) , and so on.

The *ancestral lineage* of the individual $v = (t, j)$ is

$$(s, X_s^t(j))_{-\infty < s \leq t};$$

note that eventually all ancestral lineages coalesce with the immortal line.

We say that an individual $v \in \mathcal{V}$ *descends* from an individual $u \in \mathcal{V}$ (or equivalently, u is an *ancestor* of v) if u belongs to the ancestral lineage of v . The random tree spanning \mathcal{V} which is obtained in this way is the random *look-down graph*.

Let $u = (s, i)$ and $v = (s, j)$ be two individuals living at the same time s , with $i < j$. By construction the line of ascent of v is pushed whenever the line of ascent of u is pushed, hence $Y_s^t(i) < Y_s^t(j)$, and the line of ascent of v exits not later than that of u . In this sense, the *ordering of lines by contemporaneous levels* is an *ordering by persistence*. Note also that for all times $s < t$ and all levels $i \in \mathbb{N}$:

$$Y_s^t(i) = \inf\{j \in \mathbb{N} : X_s^t(j) = i\}.$$

Thus, the time when an individual's line of ascent reaches infinity marks the time at which the individual's offspring goes extinct.

Let us note in passing that the ordering by persistence is a main distinction between the version of the look-down process developed in [DK99] and its precursor introduced in [DK96]. In the latter, the order by persistence is only stochastic, that is, lines of ascent of contemporaneous individuals at lower levels are longer “in probability”. In the modified look-down process of [DK99], explained and employed in the present paper, this property holds almost surely.

Coalescent curves and fixation curves

For $t \in \mathbb{R}$ and $i \in \mathbb{N}$ the *coalescent tree* $\mathfrak{C}^t(i)$ consists of the ancestral lineages of the individuals $(t, 1), \dots, (t, i)$, i.e. $X_s^t(1), \dots, X_s^t(i)$ for $s \leq t$, whereas the *full coalescent tree* \mathfrak{C}^t is made up of the ancestral lineages of all individuals living at time t . All these lineages eventually coalesce with the immortal line. Since any pair of ancestral lineages coalesces at rate 1, $\mathfrak{C}^t(i)$ and \mathfrak{C}^t are distributed like Kingman's (finite respectively infinite) coalescent. The number of lineages remaining at time $s < t$ can be expressed as

$$C_s^t(i) := \max\{X_s^t(1), \dots, X_s^t(i)\}, \quad C_s^t := C_s^t(\infty) := \sup_{i \in \mathbb{N}} C_s^t(i). \quad (2.3)$$

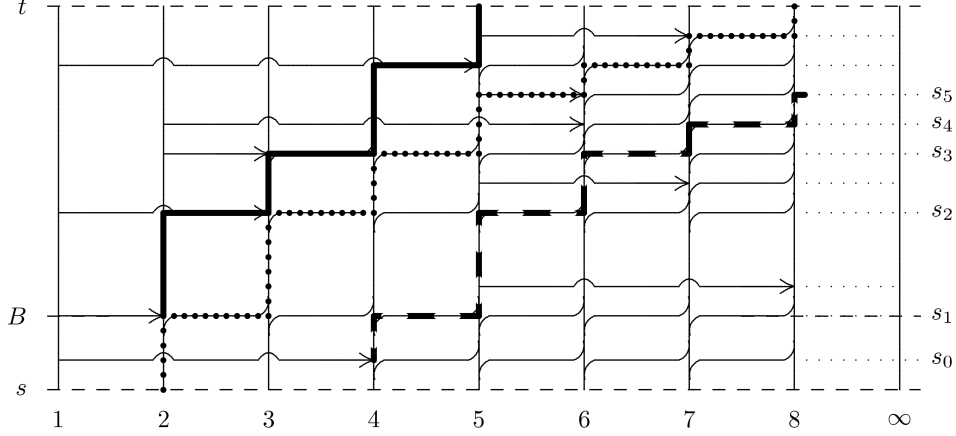


Figure 2: Detail of a look-down graph. Time is running upwards; all lines at the first 8 levels are drawn between times s and t . At times in \mathcal{P}_{ij} an arrow is drawn from i to j . All lines at levels at and above j are pushed upwards as indicated by bent lines. The solid marked line is the fixation curve $F_B^\tau, \tau \geq B$. The dotted line is the coalescent curve $C_\tau^t(8), \tau \leq t$. The dashed line is the line born at level 4 by the individual $(s_0, 1)$; it is pushed one level up at times s_1, s_2, \dots . In this picture, $X_s^t(1) = \dots = X_s^t(5) = X_s^t(7) = 1$ and $X_s^t(6) = X_s^t(8) = 2$; $Y_s^t(1) = 1, Y_s^t(2) = 6, Y_s^t(3) > 8$; $C_s^t(1) = \dots = C_s^t(5) = 1$ and $C_s^t(6) = \dots = C_s^t(8) = 2$.

In words, $C_s^t(i)$ is the number of time s -ancestors of the time t -individuals at levels $1, \dots, i$, and C_s^t is the number of time s -ancestors of the whole population at time t . For fixed t , we call $(C_s^t)_{s \leq t}$ the *coalescent curve* in the look-down graph back from time t . It is distributed like the death process in Kingman's coalescent entering from infinity.

The time when the MRCA of the total population at time t lived is

$$A_t := \sup\{s : C_s^t = 1\}.$$

All individuals at time t descend either from individual $(A_t, 1)$ or from individual $(A_t, 2)$. At time A_t a line must be born at level 2, which is equivalent to $A_t \in \mathcal{P}_{12}$. Denote the next point in \mathcal{P}_{12} after A_t by B_t :

$$B_t := \min\{s \in \mathcal{P}_{12} : s > A_t\}.$$

The offspring of the two individuals $(B_t, 1), (B_t, 2)$ evolves towards fixation in the population by pushing the line of ascent of the individual $(B_t, 3)$ towards infinity. The time E_t when this line of ascent exits equals the time when the offspring of the individual $(A_t, 2)$ is expelled by the offspring of $\{(B_t, 1), (B_t, 2)\}$. Thus the time E_t is the first time after t when a new MRCA is established, and the time when this MRCA lives is B_t .

Note that at any time τ between B_t and E_t , all the levels $1, \dots, Y_{B_t}^\tau(3) - 1$ are occupied by offspring of $\{(B_t, 1), (B_t, 2)\}$, whereas level $Y_{B_t}^\tau(3)$ is not. We therefore call

$$F_{B_t}^\tau := Y_{B_t}^\tau(3) - 1 = Y_{A_t}^\tau(2) - 1, \quad B_t \leq \tau < E_t, \quad (2.4)$$

the *fixation curve* starting in time B_t . (For the equality in (2.4), note that the line containing $(B_t, 3)$ was born at time A_t at level 2 and was pushed to level 3 at time B_t .) When $Y_{B_t}^\tau(3) = k$ the corresponding line moves to $k + 1$ at the next look-down event among the first k levels, i.e. at rate $\binom{k}{2}$. As a consequence, $F_{B_t}^\tau$ is pushed from level k to level $k + 1$ at rate $\binom{k+1}{2}$.

The MRCA point process \mathcal{F} records all the time points when the fixation curves start and end. We will pursue this in Section 4, by constructing an autonomous particle system whose trajectories give the fixation curves.

Whereas the coalescent curves are constructed from any t backwards in time, the fixation curves start only at points in \mathcal{P}_{12} and are constructed forwards in time.

At a time E when a fixation curve ends (and a new MRCA is established), all individuals descend from the MRCA who lived at the time B when this fixation curve started. Hence the fixation curve between time points B and E equals the coalescent curve back from time E . With time proceeding, the coalescent curve evolves, being more and more “zipped away” from the upper end of the fixation curve (near time point E), and still sharing the lower part (near time point B) for a while.

Having now constructed the process \mathcal{A} in terms of the look-down graph, we will study its properties in the next sections.

3 From today to the next MRCA

As in the previous section, A_t denotes the time when the current MRCA lived, E_t is the time when the next MRCA is established and $B_t = A_{E_t}$ is the time when the next MRCA lives. In this section we will compute the conditional distribution of (E_t, B_t) given A_t .

The following random variables will play a crucial role:

$$L_t := F_{B_t}^t, \quad (3.1)$$

the level at time t of the fixation curve starting at time B_t , and

$$I_t := C_{B_t}^t, \quad (3.2)$$

the level at time B_t of the coalescent curve back from time t , where we define

$$L_t := 1 \text{ and } I_t := \infty \text{ on the event } \{B_t > t\}.$$

Without loss of generality, and to ease notation, let us put $t = 0$, and write $L := L_0, I := I_0, E := E_0, B := B_0$.

Note that, because of the ordering by persistence, the lines of ascent starting at time 0 from levels $1, \dots, L$ exit only after time E , whereas the lines starting at time 0 from levels $L+1, L+2, \dots$ exit at time E or earlier. Thus, L is the random number of individuals in the present population that still have offspring when the next MRCA is established.

Proposition 3.1. *The pair (L, I) is independent of A_0 and has distribution*

$$\mathbf{P}[L = \ell, I = i] = \begin{cases} \frac{\ell - 1}{3 \binom{\ell+i}{\ell}}, & \ell \geq 2, i \geq 3 \\ \frac{1}{3}, & \ell = 1, i = \infty \\ 0, & \text{else.} \end{cases} \quad (3.3)$$

Proposition 3.1 will be proved in Section 6.

Remark 3.2. 1. Summing over i in (3.3) leads to the distribution of L :

$$\mathbf{P}[L = \ell] = \frac{2}{(\ell + 1)(\ell + 2)}, \quad \ell = 1, 2, \dots \quad (3.4)$$

Since $\{L = 1\} = \{B > 0\}$ is the event that that first fixation curve which ends after time 0 has not yet started by time $t = 0$, we infer that the probability that the next MRCA lives in today’s future is

$$\mathbf{P}[B > 0] = \mathbf{P}[L = 1] = 1/3.$$

2. Here is another quick way to (3.4), exploiting exchangeability. Recall that the number L gives the number of lines that still have offspring at the time when the next MRCA is established. At any time there are two oldest families in the population. The family sizes of these two oldest families, denoted by P and $1 - P$, evolve according to a Wright-Fisher

diffusion. It is well known (and can be understood from the Pólya urn scheme embedded in the genealogy; see e.g. facts about the Pólya-Eggenberger distribution in [JK77], eq. (4.1)) that, at any fixed time, say at time $t = 0$, P is uniformly distributed on $[0, 1]$. This also remains true conditioned on the event $A_0^\infty = -d$. By exchangeability, the probability that the first ℓ most persistent lines are in one and the $(\ell + 1)$ -st most persistent line is in the other family is

$$\mathbf{P}[L = \ell] = 2 \int_0^1 p^\ell (1 - p) dp = 2 \left(\frac{1}{\ell + 1} - \frac{1}{\ell + 2} \right) = \frac{2}{(\ell + 1)(\ell + 2)}. \quad (3.5)$$

To prepare for Theorem 1, we need one more bit of notation.

Definition 3.3. Let T_k be independent exponentially distributed random variables with parameter $\binom{k}{2}$, $k = 2, 3, \dots$, and

$$S_i^j = \sum_{k=i+1}^j T_k, \quad 1 \leq i < j \leq \infty.$$

For $d > 0$ and $i = 1, 2, \dots$ let $R_{i,d}$ be a random variable whose distribution equals the conditional distribution of S_i^∞ given that $S_1^\infty + S_i^\infty = d$.

The random variable S_i^j represents the time which Kingman's coalescent requires to come down from j to i . Consequently, $R_{i,d}$ refers to the random time for a coalescent to come down from infinity to i lines, given that coming down to 1 line requires exactly time d . Note also that S_i^j represents the time a fixation curve needs to be pushed from level i to level j . This can be seen because a fixation curve goes from level ℓ to level $\ell + 1$ whenever a look-down event among the first $\ell + 1$ levels occurs, i.e. with rate $\binom{\ell+1}{2}$.

We are now prepared to state Theorem 1, which together with Proposition 3.1 yields the desired conditional distribution of (E, B) given A_0 .

Theorem 1. Let L and I be as in (3.1) and (3.2). The conditional distribution of (E, B) , given $A_0 = -d$, $L = \ell$ and $I = i$ is represented by the random variables

$$\begin{cases} (S_1^2 + S_2^\infty, S_1^2) & \text{if } \ell = 1, \\ (S_\ell^\infty, -R_{i,d}) & \text{if } \ell > 1, \end{cases} \quad (3.6)$$

where (S_ℓ^j) and $R_{i,d}$ have the distribution specified in Definition 3.3, and $R_{i,d}$ and S_ℓ^∞ are independent.

The proof of Theorem 1 is given in Section 6.

Remark 3.4. 1. Combining (3.6) and (3.4) we obtain

$$\mathbf{P}[E \in ds | A_0 = d] = \sum_{\ell=1}^{\infty} \frac{2}{(\ell + 1)(\ell + 2)} \mathbf{P}[S_\ell^\infty \in ds], \quad s \geq 0. \quad (3.7)$$

From this one can conclude that the conditional distribution of E given A_0 is standard exponential. Indeed, think of a 2-sample (i.e. a subsample of size two) embedded in a full coalescent. The coalescence time of this 2-sample is standard exponentially distributed. Denoting by L' the number of lineages remaining in the full coalescent at the time when the 2-sample has found its common ancestor, one sees from [GT03], eq. (2.10) or [STW84], Lemma 3, or by direct calculation, that L' has the same distribution as L specified in (3.4) and (3.5). This shows that the r.h.s. of (3.7) is a decomposition of the standard exponential distribution.

2. Here is another quick (though slightly informal) argument that the waiting time to the next jump of the MRCA is exponential, independently of the depth of the current MRCA. Note first that, conditioned on $A_0 = -d$ the split of the population size into the two oldest families at time $t = 0$ is uniformly distributed on $[0, 1]$. As a consequence, given the MRCA does not jump during the time interval $[0, s]$, the split remains uniformly distributed also at time s . (This corresponds to the fact that the uniform distribution is a quasi-equilibrium for the Wright-Fisher diffusion.) At the next jump of the MRCA process one of the two oldest families dies out. After the jump there will be two families inside the surviving family that again make up a uniform split. This implies that the time between jumps proceeds in a memoryless manner, showing that the conditional distribution of E_0 given A_0 is exponential. Notably, the fact of exponential waiting times between the jumps can also be read from (3.10) in [Wat82a]. See Section 5 for comments relating to this paper and to other applications.

4 A particle representation of the MRCA point process

The set \mathcal{G} of lines defined in Section 2 randomly partitions the set $\mathcal{V} = \mathbb{R} \times \mathbb{N}$. Let us write

$$\mathcal{G}_2 := \{G \in \mathcal{G} \mid G \text{ is born at level } 2\}.$$

For each line $G \in \mathcal{G}_2$ we write $B := B(G)$ for the time when $G \in \mathcal{G}_2$ is pushed from level 2 to level 3 (due to the birth of the next line in \mathcal{G}_2) and $E := E(G)$ for the exit time of G . Thus we obtain a one-to-one correspondence between \mathcal{G}_2 and the sequence of fixation curves by associating with any $G \in \mathcal{G}_2$ the fixation curve F_B starting at time B and ending at time E . This fixation curve is related to the level path of G by $F_B^\tau = Y_B^\tau(3) - 1$ for $B \leq \tau < E$; see (2.4). The MRCA point process \mathcal{F} then can be written as

$$\mathcal{F} = \{(E, B) \mid G \in \mathcal{G}_2\}.$$

Additionally, we write

$$\eta := \{E \mid (E, B) \in \mathcal{F}\} \quad \text{and} \quad \eta_t := \eta \cap (-\infty, t]$$

for the exit time point process and its restriction to $(-\infty, t]$ respectively.

In this section we will gain more information about the processes \mathcal{F} and η by interpreting the fixation curves as the trajectories of an interacting particle system on $\{2, 3, 4, \dots\}$ whose dynamics and equilibrium distribution we will compute.

Let

$$Z_t := \#\{(E, B) \in \mathcal{F} \mid E > t, B < t\}.$$

In other words, Z_t is the number of fixation curves present at time t , that is, the number of MRCA which will be established after time t and have lived before time t .

Write

$$L_t^1 > L_t^2 > \dots > L_t^{Z_t} > 1 \tag{4.1}$$

for the levels of the fixation curves at time t . Let us interpret $(L_t^1, L_t^2, \dots, L_t^{Z_t})$ as a configuration of particles on the set of levels $\{2, 3, 4, \dots\}$ at time t , and put

$$\Lambda_t := (L_t^1, L_t^2, \dots), \quad \Lambda := (\Lambda_t),$$

where $L_t^j := 1$ for $j > Z_t$. The first components in the MRCA point process \mathcal{F} are the exit times of the “leading particles”, i.e. those time points E where $\lim_{t \uparrow E} L_t^1 = \infty$. Whenever a particle exits, the indices of all remaining particles are shifted down by one:

$$(L_E^1, L_E^2, \dots) := (L_{E-}^2, L_{E-}^3, \dots). \tag{4.2}$$

Here is a verbal description of the dynamics of the particle system (see Proposition 7.1 for a formal statement): Particles are pushed in at level 2 at rate 1, each particle at level $\ell \geq 2$ is pushed one

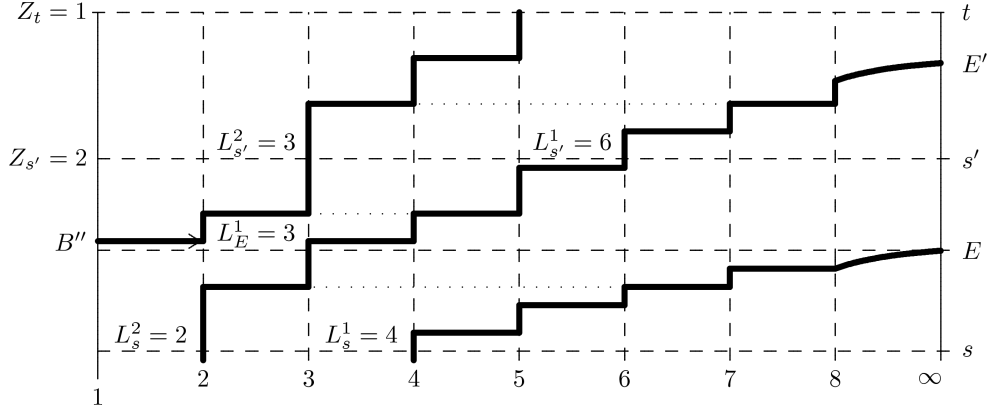


Figure 3: The embedding of the fixation curves in the look-down process. At times B fixation curves start and at times E they end. In this example, at time s the number of particles in the system is $Z_{s'} = 2$, the leading particle being at level 6, and the second particle at level 2.

level up at rate $\binom{\ell+1}{2}$, and this is done in a coupled way such that, whenever a particle is pushed, all particles at higher levels are pushed simultaneously. The next theorem specifies the equilibrium distribution of Λ . We will see that this distribution prevails also in the distinguished random time points E where $\lim_{t \uparrow E} L_t^1 = \infty$. This property is crucial to see that η is a Poisson process.

Theorem 2. 1. *The process $\Lambda = (\Lambda_t)$ is Markov with stationary distribution*

$$\pi_{\Lambda}(\ell_1, \ell_2, \dots) = \begin{cases} \frac{1}{3} \prod_{j: \ell_j > 1} \frac{2}{(\ell_j + 2)(\ell_j - 1)}, & \ell_k > \ell_{k+1}, \quad \text{if } \ell_k > 1, \\ 0, & \text{else.} \end{cases} \quad (4.3)$$

In particular, the stationary distribution of L^1 is

$$\pi_{L^1}(\ell) = \frac{2}{(\ell + 1)(\ell + 2)}. \quad (4.4)$$

2. *The process of exit times η is a stationary Poisson process.*

Remark 4.1. 1. The “arrival time points” B of the particles in the system Λ (the times when the MRCAs live) are the points of the stationary Poisson process \mathcal{P}_{12} . Theorem 2 states that also the “departure time points” E (the times when the MRCAs are established) form a stationary Poisson process. Thus, the Poisson input process of times B when the fixation curves begin is transformed by a “dependent stochastic shift” into the Poisson output process of times E when they end. This is similar to Burke’s theorem which states that the departure process in a time stationary $M/M/1$ queue is Poisson, see [Kur98] and references given there. A crucial property (proved already by Burke (1956)) is that in a stationary $M/M/1$ queue the distribution of the queue length at time t is independent of the departure times $\leq t$. In the language of queueing theory, our particles correspond to customers entering at the time points of a stationary Poisson process, and the time which a typical customer spends in the system is distributed like S_2^∞ specified in Definition 3.3. These times are mutually dependent. As the proof of Theorem 2 reveals, like in Burke’s theorem the state of the system (now given by the configuration Λ_t of particles at time t) does not depend on the departure times $\leq t$.

2. We recently learned from Tom Kurtz about the manuscript [DK06] where he and Peter Donnelly have established the filtered martingale problem for the N -level analogue of the

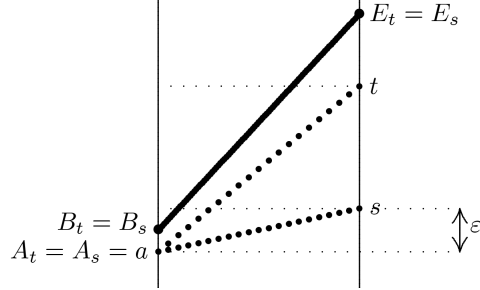


Figure 4: Assume we know $A_t = A_s = a$ for some $s = a + \varepsilon$. This knowledge leads to a higher chance of the MRCA time B_t falling between times a and s than in an equilibrium situation. This shows that the future of the process \mathcal{A} at time t depends on the past and \mathcal{A} cannot be Markov. See text for explanation.

pair (Λ, η) in the context of [Kur98], Theorem 3.2 and thus achieved an alternative proof of the fact that the “MRCA fixation process” η is Poisson.

3. Whereas the particle process Λ is Markov, the MRCA process \mathcal{A} is not. This can be seen as follows:

Let a, s, t be as in Figure 3. Conditioned on $A_t = a$ we obtain from Theorem 1:

$$\mathbf{P}[B_t < s | A_t = a] = \mathbf{P}[R_{I, t-a} > t - s] \xrightarrow{s \downarrow a} 0.$$

On the other hand, we claim that $\mathbf{P}[B_t < s | A_t = A_s = a]$ does not converge to zero as $s \downarrow a$, which shows that \mathcal{A} cannot be Markov. To verify the claim, we write, using Bayes’ rule

$$\mathbf{P}[B_t < s | A_t = A_s = a] = \frac{\mathbf{P}[B_s < s, A_t = A_s | A_s = a]}{\mathbf{P}[A_t = A_s | A_s = a]}.$$

By Theorem 1, the denominator converges to $e^{-(t-a)}$ as $s \downarrow a$. Likewise, the numerator is bounded away from 0 as $s \downarrow a$, a trivial lower bound being

$$\mathbf{P}[L_s = 2] \mathbf{P}[S_2^3 \geq t - s] = \frac{1}{6} e^{-3(t-s)} \geq \frac{1}{6} e^{-3(t-a)}.$$

4. The level L_t^1 of the leading particle in (4.1) coincides with L_t defined in (3.1). Thus we recover (3.4) from (4.4).

Recall from (4.1) that

$$Z_t = \max\{j \in \mathbb{N} \mid L_t^j > 1\},$$

where $\max \emptyset := 0$. Consequently,

$$\{Z_t = 0\} = \{L_t^1 = 1\}.$$

This is the event that there is no particle on $\{2, 3, 4, \dots\}$ at time t , or equivalently, that all fixation curves starting before time t also end before t . Given this event, the fixation curves starting before time t are independent of those starting after t .

In a way, the random variable Z_t of MRCA that are established in today’s future and live in today’s past measures the dependence between past and future in the MRCA process. Note also that because of Theorem 3 the distribution of Z_t does not change when t is conditioned to be the time of an MRCA change.

In the next theorem we calculate the equilibrium distribution of $Z := Z_t$.

Theorem 3. 1. The probability generating function of Z is

$$\mathbf{E}[u^Z] = \frac{1}{3} \exp \left(\sum_{i=2}^{\infty} \log \left(\frac{i(i+1) + 2(u-1)}{(i+2)(i-1)} \right) \right).$$

2. The expectation and variance are

$$\mathbf{E}[Z] = 1, \quad \mathbf{Var}[Z] = 14 - \frac{4}{3}\pi^2 \approx 0.84052.$$

3. The probability weights are given by

$$\mathbf{P}[Z = z] = \frac{2^z}{3} \sum_{\mathbf{a}: \sum i a_i = z} (-1)^{z+\sum a_i} \frac{1}{a_1! \cdots a_z!} \prod_{j=1}^z \left(\frac{x_j}{j}\right)^{a_j}, \quad (z \geq 0)$$

where the $a_i \in \mathbb{N}_0$ and

$$\begin{aligned} x_k &:= \frac{(-1)^{k+1}}{3^{2k-1}} \sum_{j=1}^k \binom{2k-j-1}{k-j} 3^{j-1} (b_j - 1_{\{j \text{ even}\}} 2\zeta(j)), \\ \zeta(k) &:= \sum_{j=1}^{\infty} \frac{1}{j^k}, \\ b_j &:= 1 + \frac{1}{2j} + \frac{1}{3j}. \end{aligned} \tag{4.5}$$

4. The weights for $z = 0, 1, 2, 3$ are

$$\begin{aligned} \mathbf{P}[Z = 0] &= \frac{1}{3}, & \mathbf{P}[Z = 1] &= \frac{11}{27} \approx 0.40740, \\ \mathbf{P}[Z = 2] &= \frac{107}{243} - \frac{2}{81}\pi^2 \approx 0.19664, & \mathbf{P}[Z = 3] &= \frac{1003}{2187} - \frac{10}{243}\pi^2 \approx 0.05246. \end{aligned}$$

5 Relations to population genetics

Consider sequence data, obtained from a sample of individuals in a population that reproduces according to Wright-Fisher dynamics. Besides resampling we consider neutral mutations for an infinite sites model (as introduced in [Kim71]) occurring at rate $\theta/2$ along each line. Using common notation in population genetics, we consider the diffusion limit of the dynamics of the population, where time has been rescaled by a factor N , the number of haploids in the population. The per generation mutation probability of μ along each line is rescaled to $\theta/2$, where $\theta = 2N\mu$.

Mutations can also be modelled in the look-down picture: for each level there is an independent Poisson clock with rate $\theta/2$ by which mutations on the line carrying the corresponding level accumulate. This implies that on each line of the lookdown process mutations arise at rate $\theta/2$. All mutations an individual carries at time t are collected along its line of descent.

Segregating sites

For two individuals sampled from the population, the expected number of segregating sites is $\theta \mathbf{E}[T_c]$, where T_c is the random time to coalescence of the individuals' ancestral lineages. This time is unusually short at instances when the MRCA changes. In [Taj90], Tajima studied the coalescent at such times. He concluded that then the coalescence rate from k to $k-1$ ancestral lineages is $\binom{k+1}{2}$, his argument being that, in addition to the k lineages, there is one extra line, which apparently must belong to the family that disappears at the time of the MRCA change.

These coalescence rates can also be seen from the particle representation of the MRCA process. In fact, the fixation curves give the shape of the coalescent tree of the whole population back from the time of the MRCA change. Recall that the fixation curve moves from level k to $k+1$ at rate $\binom{k+1}{2}$. Consequently the time the coalescent back from some time point E stays with k lines is exponentially distributed with rate $\binom{k+1}{2}$, which means that this is the rate to go down from k to $k-1$ lineages. As these rates differ from the rates in Kingman's coalescent the random coalescence time T_c of the 2-sample cannot be exponential. However, the 2-sample coalescent is embedded in

the full coalescent; the probability that the two sampled lines find a common ancestor at the time when there are ℓ lines left in the full coalescent is (see Remark 3.4 or [Taj90], equation (6))

$$\frac{2}{(\ell+1)(\ell+2)}.$$

As the time of going down from infinity to ℓ lines in the coalescent at an MRCA time is distributed like $S_{\ell+1}^\infty$, we obtain the distribution for the coalescence time T_c of the two lines

$$\mathbf{P}[T_c \in dt] = \sum_{\ell=1}^{\infty} \frac{2}{(\ell+1)(\ell+2)} \mathbf{P}[S_{\ell+1}^\infty \in dt].$$

Taking expectations we obtain

$$\mathbf{E}[T_c] = \sum_{\ell=1}^{\infty} \frac{2}{(\ell+1)(\ell+2)} \frac{2}{\ell+1} = 4 \sum_{\ell=1}^{\infty} \left(\frac{1}{(\ell+1)^2} - \frac{1}{(\ell+1)(\ell+2)} \right) = \frac{2}{3}\pi^2 - 6 \approx 0.58,$$

a result already obtained in [Taj90]. As the coalescence time for two lines in equilibrium is exponential with mean 1, this result means that the expected number of segregating sites for a 2-sample is reduced by 42% at times when the MRCA changes.

For samples of arbitrary size, the number of segregating sites is Poisson with mean $\theta/2$ times the total branch length of the sample's genealogical tree. In [RBY04], Figure 2c, a path of the time evolution of this total branch length is depicted for a spatial and a “well-mixed” population. At certain instances, one sees sudden substantial decrease of the path length. One may guess that this happens primarily at times at which the MRCA changes, since then the coalescent tree is unusually short.

Substitutions

Most mutations that occur in a population are quickly lost. However, some eventually fixate, i.e. all individuals in the population carry the new mutation. This replacement is termed a *substitution* and the corresponding mutations are called *determining mutations*. In [Wat82a] and [Wat82b], Watterson studied several aspects of the process of substitutions. While we are concerned with the jump from today's MRCA to the next one, Watterson fixes two time points 0 and t and studies the time between the MRCAs at these times, i.e. $A_t - A_0$, irrespectively of the number of MRCAs that are established between 0 and t . All mutations on the ancestral line between A_t and A_0 are then determining mutations and their number gives the the number of substitutions between times 0 and t .

The only way a mutation can become a substitution is through an MRCA change. This is because any mutation that occurs in the population belongs to one of the two oldest families. For the mutation to become fixed it is necessary that the family not carrying the mutation dies out. In other words, it is necessary that the MRCA changes.

Consider the graphical lookdown representation including mutations falling on lines at all levels at rate $\theta/2$. A mutation that occurs is determining if and only if it occurs on the line at level one. Indeed, we already found that MRCAs of the population as seen in the lookdown picture always are at level one. On the other hand, given the time point of a mutation on a line at level one, eventually all individuals in the population are descendants of the individual carrying this mutation which shows that all mutations that occur at level one are determining.

Denote by $\mathcal{S} = \{(\tilde{E}, S)\}$ the process of times $\{\tilde{E}\}$ and number $\{S\}$ of substitutions at these times. As times E of MRCA changes are the only ones that can be substitution times and the number of mutations on a line is Poisson distributed with rate $\frac{\theta}{2}$ we find that the process \mathcal{S} is a close relative to the MRCA point process:

Proposition 5.1. *Let $\{(E, B)\}$ be distributed as the MRCA point process. Additionally, for all successive pairs (E', B') and (E'', B'') , let S'' be Poisson-distributed with intensity parameter $\frac{\theta}{2}(B'' - B')$. Then $\{(E'', S'') : S'' > 0\}$ is a version of \mathcal{S} .*

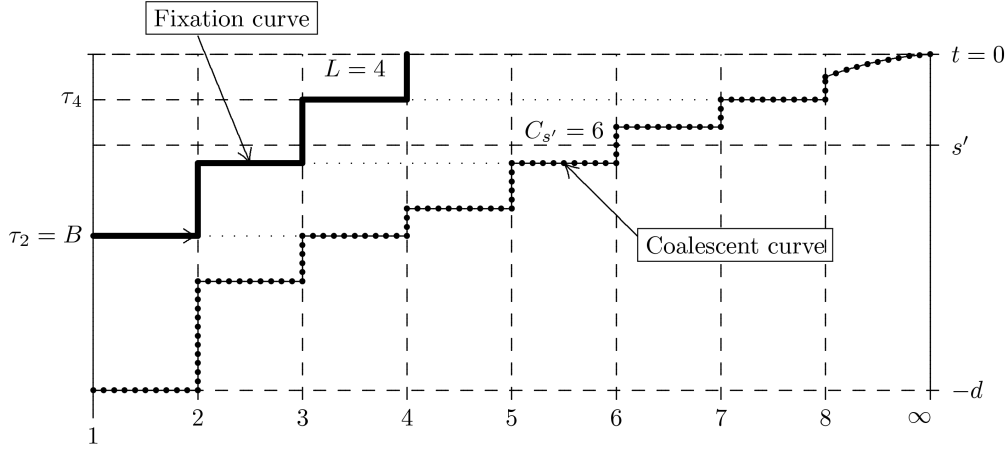


Figure 5: The variable I^k is the level of the coalescence curve back from time $t = 0$ when the next fixation curve has reached level k . The corresponding real times are denoted by τ_k . The variable K^j is the level of the fixation curve when the coalescent curve has reached level j . In this example, $I^2 = 4$, $I^3 = 6$, $I^4 = 8$, $K^2 = K^3 = 1$, $K^4 = K^5 = 2$, $K^6 = K^7 = 3$ and $K^8 = 4$.

This confirms the observation in [Wat82b] that (i) substitution times do not form a Poisson process and (ii) substitutions tend to occur in clusters.

6 Proof of Theorem 1

Recall the definition of L_t and I_t in (3.1) and (3.2), and also recall that we put without loss of generality $t = 0$, omitting the corresponding sub- and superscripts 0. By definition, the fixation curve F_B starts at time B at level 2 and exits at time E at level ∞ ; let us now extend this definition by putting

$$F_B^\tau := 1 \quad \text{if } \tau < B.$$

The following auxiliary variables will be helpful:

$$K^j := \text{the level of } F_B \text{ while } C = j, \quad j = 2, 3, \dots$$

and

$$I^k := \text{the level of } C \text{ when } F_B \text{ reaches level } k, \quad k = 2, 3, \dots$$

Formally, putting

$$\tau_k := \inf\{\tau : F_B^\tau = k\}, \quad k = 2, 3, \dots$$

we have

$$I^k = \begin{cases} C_{\tau_k}, & \text{if } \tau_k < 0, \\ \infty, & \text{if } \tau_k \geq 0. \end{cases}$$

Thus, $\tau_2 = B$, $I^2 = I$,

$$K^j = \max\{k : I^k \leq j\} \quad \text{and} \quad L = K^\infty := \lim_{j \rightarrow \infty} K^j. \quad (6.1)$$

The random variables I^k , K^j and L are illustrated in Figure 5.

Lemma 6.1. $K = (K^2, K^3, \dots)$ is an inhomogeneous Markov chain starting in $K^2 = 1$ and with transition probability given by

$$\mathbf{P}[K^{j+1} = k+1 | K^j = k] = \frac{\binom{k+1}{2}}{\binom{j+1}{2}} = 1 - \mathbf{P}[K^{j+1} = k | K^j = k], \quad j > k \geq 1. \quad (6.2)$$

Moreover, K is independent of the coalescence curve C .

Proof. When the coalescence curve moves from $j+1$ to j at some time s , that is, $C_s = j+1$ and $C_{s-} = j$, then some look-down event involving two levels $\leq k+1$ must happen. When at time $s-$ the fixation curve is at level k , the probability that the fixation curve jumps at time s from k to $k+1$ is $\binom{k+1}{2}/\binom{j+1}{2}$, since all possible look-down events are equally probable and there are $\binom{\ell+1}{2}$ events that push the next fixation curve one level up. Observe that this is independent of the prehistory K^2, \dots, K^{j-1} , independent of the coalescence curve C , and in particular also independent of the time A_0 . \square

In the next lemma we calculate the joint distribution of the random variables I^k .

Lemma 6.2. The joint distribution of (I^2, I^3, \dots) is given by

$$\mathbf{P}[I^2 = \dots = \infty] = \frac{1}{3}, \quad (6.3)$$

$$\mathbf{P}[I^2 = i_2, \dots, I^\ell = i_\ell, I^{\ell+1} = \dots = \infty] = \frac{\ell!(\ell-1)!}{3} \prod_{m=2}^{\ell} \frac{1}{(i_m + m)(i_m + m - 1)} \quad (6.4)$$

for $2 < i_2 < \dots < i_\ell$.

Proof. The event $\{I^2 = \infty\}$ equals the event that the next fixation curve has not yet started by time 0, that is the event $\{B > 0\} = \{K_j = 1 \text{ for all } j = 1, 2, \dots\}$. Thus, using (6.2),

$$\mathbf{P}[I^2 = \infty] = \prod_{j=3}^{\infty} \left(1 - \frac{1}{\binom{j}{2}}\right) = \prod_{j=3}^{\infty} \frac{(j+1)(j-2)}{j(j-1)} = \frac{1}{3},$$

since the product telescopes. This shows (6.3). To prove (6.4), we express the event on its left hand side in terms of the variables K^j :

$$\begin{aligned} & \{I^2 = i_2, \dots, I^\ell = i_\ell, I^{\ell+1} = \dots = \infty\} \\ &= \{K^2 = \dots = K^{i_2-1} = 1, \dots, K^{i_{\ell-1}} = \dots = K^{i_\ell-1} = \ell-1, K^{i_\ell} = K^{i_{\ell+1}} = \dots = \ell\}. \end{aligned}$$

Putting $i_1 = 2$ and $i_{\ell+1} = \infty$, and using Lemma 6.1 we arrive at

$$\begin{aligned} \mathbf{P}[I^2 = i_2, \dots, I^\ell = i_\ell, I^{\ell+1} = \infty] &= \left[\prod_{m=1}^{\ell} \prod_{j=i_m+1}^{i_{m+1}-1} \left(1 - \frac{\binom{m+1}{2}}{\binom{j}{2}}\right) \right] \cdot \left[\prod_{m=1}^{\ell-1} \frac{\binom{m+1}{2}}{\binom{i_{m+1}}{2}} \right] \\ &= \ell!(\ell-1)! \left[\prod_{m=1}^{\ell} \prod_{j=i_m+1}^{i_{m+1}-1} \frac{(j-m-1)(j+m)}{j(j-1)} \right] \left[\prod_{m=1}^{\ell-1} \frac{1}{(i_{m+1})(i_{m+1}-1)} \right] \\ &= \ell!(\ell-1)! \left[\prod_{m=1}^{\ell-1} \frac{(i_m - m) \cdots (i_m - 1)}{(i_{m+1} - m - 1) \cdots (i_{m+1} - 2)(i_{m+1} - 1)} \right. \\ &\quad \left. \frac{i_{m+1} \cdots (i_{m+1} + m - 1)}{(i_m + 1) \cdots (i_m + m)i_{m+1}} \right] \cdot \frac{(i_\ell - \ell) \cdots (i_\ell - 1)}{(i_\ell + 1) \cdots (i_\ell + \ell)} \\ &= \frac{\ell!(\ell-1)!}{3} \frac{1}{(i_\ell + \ell)(i_\ell + \ell - 1)} \left[\prod_{m=2}^{\ell-1} \frac{1}{(i_m + m)(i_m + m - 1)} \right] \\ &= \frac{\ell!(\ell-1)!}{3} \prod_{m=2}^{\ell} \frac{1}{(i_m + m)(i_m + m - 1)} \end{aligned}$$

\square

From the joint distribution of I^2, I^3, \dots given in Lemma 6.2 we obtain because of (6.1) the joint distribution of (L, I) by projection:

Proof of Proposition 3.1

Because of $\{L = 1\} = \{I^2 = \infty\}$, we obtain the assertion of (3.3) for $\ell = 1$ from (6.3). For $\ell = 2, 3, \dots$ we proceed by induction. For $\ell = 2$ we have, using again Lemma 6.2,

$$\mathbf{P}[I = i, L = 2] = \mathbf{P}[I^2 = i, I^3 = \infty] = \frac{2}{3} \frac{1}{(i+2)(i+1)}.$$

If the assertion is true for all $2, \dots, \ell$, we have

$$\begin{aligned} \mathbf{P}[I = i, L = \ell + 1] &= \sum_{i < i_3 < \dots < i_{\ell+1}} \mathbf{P}[I^2 = i, I^3 = i_3, \dots, I^{\ell+1} = \ell + 1, I^{\ell+2} = \infty] \\ &= \frac{1}{(i+2)(i+1)} \sum_{i < i_3 < \dots < i_{\ell+1}} \frac{(\ell+1)!}{3} \prod_{m=3}^{\ell+1} \frac{1}{(i_m + m)(i_m + m - 1)} \\ &= \frac{(\ell+1)\ell}{(i+2)(i+1)} \sum_{i < j} \frac{1}{(j+3)(j+2)} \sum_{j+1 < i_3 < \dots < i_{\ell}} \frac{\ell!(\ell-1)!}{3} \prod_{m=3}^{\ell} \frac{1}{(i_m + m)(i_m + m - 1)} \\ &= \frac{(\ell+1)\ell}{(i+2)(i+1)} \sum_{i < j} \mathbf{P}[I = j+1, L = \ell] = \frac{(\ell+1)\ell(\ell-1)}{3(i+2)(i+1)} \sum_{i < j} \frac{1}{(j+2) \cdots (j+\ell+1)} \\ &= \frac{(\ell+1)\ell}{3(i+2)(i+1)} \sum_{i < j} \frac{1}{(j+2) \cdots (j+\ell)} - \frac{1}{(j+3) \cdots (j+\ell+1)} \\ &= \frac{(\ell+1)\ell}{3(i+2)(i+1)} \frac{1}{(i+3) \cdots (i+\ell+1)} \end{aligned}$$

and we are done. \square

We turn now to the

Completion of the Proof of Theorem 1

Given $\{L = 1\} = \{B > 0\}$, and independently of C (and therefore also of A_0), the time B it takes to enter the next fixation curve is standard exponentially distributed (and therefore distributed like $S_1^2 \sim \exp(1)$), and the additional time it takes this fixation curve to exit is distributed like S_2^∞ , and is independent of B .

Given $L = \ell \geq 2$, $I = i < \infty$ and $A_0 = -d$, the time at which the coalescent curve jumps from level $i+1$ to i is distributed like $-R_{i,d}$. By construction, this is also the time B at which the next fixation curve F_B enters. At time 0, this fixation curve is at level ℓ ; independently of the past, the time it takes until this fixation curve exits is distributed like S_ℓ^∞ . \square

7 Proof of Theorem 2

First we give a formal description of the dynamics of the process Λ . Afterwards we derive its equilibrium distribution, and finally we show that this equilibrium distribution also prevails at the distinguished times E .

The dynamics of Λ

Assume $Z_t = k$ with $L_t^1 = \ell_1, \dots, L_t^k = \ell_k > L_t^{k+1} = 1$, i.e. at time t there are exactly k particles at levels > 1 ; in other words, exactly k fixation curves are present at time t . Assume level j looks down to level i for $i < j$. If $j > \ell_1 + 1$ only lines at level greater than $\ell_1 + 1$ are pushed. In this case

no particle moves, i.e. L_t stays constant. When $j \leq \ell_1 + 1$, at least the level of the next fixation curve increases by one from ℓ_1 to $\ell_1 + 1$ and the corresponding particle moves. The rate of these events is $\binom{\ell_1+1}{2}$ which equals the rate at which a fixation curve moves from ℓ_1 to $\ell_1 + 1$. When j is at most $\ell_2 + 1$, also the position of the second fixation curve is increased and the corresponding particle moves.

As look-down events among the first $\ell_1 + 1$ levels occur at rate $\binom{\ell_1+1}{2}$, this is also the rate at which the first particle moves. To be exact, with rate $\binom{\ell_1+1}{2} - \binom{\ell_2+1}{2}$ only the first particle is affected, with rate $\binom{\ell_2+1}{2} - \binom{\ell_3+1}{2}$ the first two particle move and so on. Additionally, at rate 1, a look-down event from level 2 to 1 occurs which has the effect that a new particle enters at level 2, i.e. and L_t^{k+1} moves from level 1 to level 2 and all particles at levels greater than 1 move as well.

The first particle moves at a quadratic rate and thus reaches infinity within finite time. When it hits infinity at time E the fixation curve is completed and $L_E^k = L_{E-}^{k+1}$ for $k \geq 1$ as stated in (4.2) because at time E the second particle becomes the leading one.

The just stated arguments prove the following proposition describing the dynamics of the process Λ .

Proposition 7.1. *From $\Lambda_t = (\ell_1, \ell_2, \dots)$, transitions occur*

$$\begin{cases} \text{to } (\ell_1 + 1, \dots, \ell_k + 1, \ell_{k+1}, \dots) \text{ at rate } \binom{\ell_k+1}{2} - \binom{\ell_{k+1}+1}{2} & \text{if } \ell_k > 1, \\ \text{to } (\ell_1 + 1, \dots, \ell_k + 1, \ell_{k+1}, \dots) \text{ at rate } 1 & \text{if } \ell_{k-1} > \ell_k = 1. \end{cases}$$

To derive the equilibrium distribution of the particle system Λ , it will be helpful to compute the one-time distributions and the limiting distribution of the Markov chain K from Lemma 6.1.

Lemma 7.2.

$$\mathbf{P}[K^j = k] = \frac{j+1}{j-1} \frac{2}{(k+1)(k+2)}, \quad j > k \geq 1. \quad (7.1)$$

Proof. To prove (7.1), we will proceed by induction. Because of $\mathbf{P}[K^2 = 1] = 1$, the formula is true for $j = 2$. From (6.2) we obtain the induction step:

$$\begin{aligned} \mathbf{P}[K^{j+1} = k] &= \mathbf{P}[K^j = k] \left(1 - \frac{\binom{k+1}{2}}{\binom{j+1}{2}}\right) + \mathbf{P}[K^j = k-1] \frac{\binom{k}{2}}{\binom{j+1}{2}} \\ &= \frac{j+1}{j-1} \frac{2}{(k+1)(k+2)} \frac{(j+1)j - (k+1)k}{(j+1)j} + \frac{j+1}{j-1} \frac{2}{k(k+1)} \frac{k(k-1)}{(j+1)j} \\ &= \frac{1}{(j-1)j} \frac{2}{(k+1)(k+2)} ((j+1)j - (k+1)k + (k-1)(k+2)) \\ &= \frac{j+2}{j} \frac{2}{(k+1)(k+2)}. \end{aligned}$$

□

We are now ready for the

Completion of the Proof of Theorem 2

We will briefly write $(L^1, L^2, \dots) := (L_0^1, L_0^2, \dots)$. Observe that L^1 equals the level L which the fixation line entering at time B has reached at time 0. From (6.1) and Lemma 7.2 we thus infer readily that

$$\mathbf{P}[L_1 = \ell_1] = \frac{2}{(\ell_1 + 1)(\ell_1 + 2)}, \quad (7.2)$$

which proves (4.4) and also re-establishes (3.4).

Next we compute the conditional distribution of L^{k+1} , given $L^k = \ell_k, \dots, L^1 = \ell_1$, where $2 \leq \ell_k < \dots < \ell_1$. Consider the k -th particle, i.e. the particle which has level ℓ_k at time 0,

and denote the time at which this particle entered at level 2 by B_k . Since the trajectory of this particle between times B_k and 0 is an initial piece of the coalescent curve (belonging to the exit time of this particle), and since L^{k+1} is the level of the next fixation curve while this coalescent curve has level ℓ_k , we can apply Lemmata 6.1 and 7.2 to the trajectory of the $(k+1)$ -st particle, parametrised by the levels of the k -th particle's trajectory, to conclude that

$$\mathbf{P}[L^{k+1} = \ell_{k+1} | L^k = \ell_k, \dots, L^1 = \ell_1] = \frac{\ell_k + 1}{\ell_k - 1} \frac{2}{(\ell_{k+1} + 1)(\ell_{k+1} + 2)}. \quad (7.3)$$

Iterating this we obtain

$$\begin{aligned} & \mathbf{P}[L^1 = \ell_1, \dots, L^k = \ell_k, L^{k+1} = 1] \\ &= \frac{2}{(\ell_1 + 1)(\ell_1 + 2)} \frac{\ell_1 + 1}{\ell_1 - 1} \frac{2}{(\ell_2 + 1)(\ell_2 + 2)} \frac{\ell_2 + 1}{\ell_2 - 1} \cdots \frac{2}{(\ell_k + 1)(\ell_k + 2)} \frac{\ell_k + 1}{\ell_k - 1} \frac{1}{3} \\ &= \frac{1}{3} \prod_{j=1}^k \frac{2}{(\ell_j + 2)(\ell_j - 1)}. \end{aligned}$$

This shows (4.3).

In Section 6 we argued, by disentangling the combinatorics from the time embedding, that $L = L_1$ is independent of the coalescence curve $C = C^0$. The same argument shows that Λ_t is independent of (C^t, η_t) , that is, both the coalescent curve back from time t and the exit time points before t .

We claim that this assertion remains true conditioned on $\{t \in \eta\}$, i.e. the event that t is an exit time. Indeed, given $\{t \in \eta\}$ we know that t is the exit point of a fixation curve, which hence must coincide with the coalescence curve C^t . So the above argument shows that also under this additional conditioning the particle configuration Λ_t is independent of C^t and η_t .

Now we turn to assertion 2. of the theorem. Consider a population in equilibrium. We know already that Λ_t is in equilibrium, i.e. has distribution π_Λ , independently of the exit times of particles before t and no matter if t is conditioned to be an exit time or not. This proves that η is a stationary renewal process. Additionally we know that waiting times between points have the same distribution as the waiting time out of equilibrium. Thus the waiting times are memoryless, hence exponential, and η is Poisson. \square

Remark 7.3. Here is a more heuristic way (in the spirit of Remark 3.2.2) to see the identity

$$\mathbf{P}[L^2 = \ell_2 | L^1 = \ell_1] = \frac{\ell_1 + 1}{\ell_1 - 1} \frac{2}{(\ell_2 + 1)(\ell_2 + 2)}. \quad (7.4)$$

Equation (3.5) says that $L = L_1$ has the distribution of the initial run length R in a coin tossing with random, uniformly on $[0, 1]$ distributed success probability. Similarly, equation (7.3) says that

$$\text{given } L_1 = \ell_1, \text{ the random variable } L_2 \text{ is distributed like } R \text{ conditioned to } \{R < \ell_1\}. \quad (7.5)$$

This is readily seen because

$$\mathbf{P}[R < \ell] = 1 - 2 \int_0^1 p^\ell dp = \frac{\ell - 1}{\ell + 1},$$

and consequently

$$\mathbf{P}[R = \ell_2 | R < \ell_1] = \frac{\ell_1 + 1}{\ell_1 - 1} \frac{2}{(\ell_2 + 1)(\ell_2 + 2)}.$$

The property (7.5) can also be understood as follows: $L^1 = \ell_1$ is the number of currently living individuals that still have offspring at the time E_0 of the next MRCA change and L^2 is the number

of individuals still having offspring at the time E_{E_0} of the next but one MRCA change. At time E_0 one of the two families which were the oldest at time 0, dies out, and our condition is that the ℓ_1 individuals at time 0 have offspring in the surviving family. This surviving family will again be made up of two oldest subfamilies, whose sizes again constitute a uniform split of $[0, 1]$. At the time of the next but one MRCA change after time 0, at least *some* of the ℓ_1 lines must have gone extinct, which amounts to the condition that not all of them belong to the same subfamily. The number L^2 of lines that belong to the surviving subfamily thus has the same distribution as R conditioned to $\{R < \ell_1\}$.

8 Proof of Theorem 3

By the definition of Z we immediately see from Theorem 2 that in equilibrium

$$\mathbf{P}[Z = z] = \sum_{1=\ell_{z+1} < \ell_z < \dots < \ell_1} \pi_\Lambda(\ell_1, \ell_2, \dots) = \frac{1}{3} \sum_{1 < \ell_z < \dots < \ell_1} \prod_{j=1}^z \frac{2}{(\ell_j + 2)(\ell_j - 1)} \quad (8.1)$$

This is the basis for the proof of Theorem 3. We first show that the correct weights of the distribution of Z are given by 3. The weights from 4. are just an application of this. From the weights we compute the probability generating function given in 1. By calculating derivatives we obtain the expectation and the variance as given in 2.

Proof of 3. and 4.

All we have to do is to simplify (8.1) for more efficient computation. Therefore we define

$$\begin{aligned} f(\ell) &:= \frac{1}{(\ell + 2)(\ell - 1)}, \\ x_k &:= \sum_{\ell=2}^{\infty} (f(\ell))^k, \end{aligned} \quad (8.2)$$

(the definition of x_k matches the definition in (4.5) as we will show below) and

$$\begin{aligned} p_0 &:= 1, \\ p_z &:= \sum_{1 < \ell_z < \dots < \ell_1} \prod_{m=1}^z f(\ell_m) = \frac{1}{z!} \sum_{1 < \ell_z, \dots, \ell_1 \text{ pwd}} f(\ell_1) \cdots f(\ell_z). \quad (z > 0) \end{aligned}$$

Here *pwd* means *pairwise different*. With this definition, for $z \geq 0$,

$$\mathbf{P}[Z = z] = \frac{2^z}{3} p_z.$$

We will show first

$$p_z = \frac{1}{z} \left(\sum_{j=1}^z (-1)^{j-1} p_{z-j} x_j \right), \quad (8.3)$$

with x_j given by (4.5), which gives p_z recursively. Then we calculate p_z as

$$p_z = \sum_{\underline{a}: \sum ia_i = z} (-1)^{z + \sum a_i} \frac{1}{a_1! \cdots a_z!} \prod_{j=1}^z \left(\frac{x_j}{j} \right)^{a_j}, \quad (8.4)$$

which gives part 2. of Theorem 3.

For (8.3) define

$$b_{z,k} := \frac{1}{(z-1)!} \sum_{1 < \ell_1, \dots, \ell_z \text{ p.w.d.}} f(\ell_1) \cdots f(\ell_{z-1}) \cdot f(\ell_z)^k.$$

Then $p_z = b_{z,1}/z$, $x_k = b_{1,k}$, and consequently

$$\begin{aligned} b_{z+1,k} &= \frac{1}{z!} \sum_{1 < \ell_1, \dots, \ell_z} f(\ell_1) \cdots f(\ell_z) \left(\left(\sum_{j=2}^{\infty} f(j)^k \right) - f(\ell_1)^k - \dots - f(\ell_z)^k \right) = \frac{1}{z} b_{z,1} x_k - b_{z,k+1} \\ &= p_z x_k - b_{z,k+1}. \end{aligned}$$

Therefore we can write

$$z p_z - (-1)^{z-1} x_z = b_{z,1} + (-1)^{z-2} b_{1,z} = \sum_{j=1}^{z-1} (-1)^{j-1} (b_{z+1-j,j} + b_{z-j,j+1}) = \sum_{j=1}^{z-1} (-1)^{j-1} p_{z-j} x_j.$$

Here the second equality follows because the sum telescopes. This gives (8.3).

The second equation, (8.4) is proved by induction. Instead of (8.4) we prove

$$p_z = \sum_{k=1}^z (-1)^{z+k} \frac{1}{k!} \sum_{\underline{j}: j_1 + \dots + j_k = z} \prod_{i=1}^k \frac{x_{j_i}}{j_i} \quad (8.5)$$

which then gives (8.4) as the sum is over all vectors \underline{j} of length k which sum up to z . Every such vector can be translated into a configuration \underline{a} with $\sum i a_i = z$ where a_i is the number of i 's in \underline{j} . As for a given length k of the vector \underline{j} there are $\frac{k!}{a_1! \cdots a_k!}$ of these vectors leading to the same configuration (8.4) is the same as (8.5).

For $z = 1$ (8.5) gives $p_1 = x_1$ which is true by definition of p_z and x_z . Assume the formula is correct for $1, \dots, z$ and use (8.3) to conclude that

$$\begin{aligned} p_{z+1} &= \frac{1}{z+1} \left(\sum_{j=1}^z (-1)^{j-1} x_j \sum_{k=1}^{z+1-j} (-1)^{z+1-j+k} \frac{1}{k!} \sum_{\underline{j}: j_1 + \dots + j_k = z+1-j} \prod_{i=1}^k \frac{x_{j_i}}{j_i} \right) + \frac{(-1)^z}{z+1} x_{z+1} \\ &= \left(\sum_{k=1}^z (-1)^{z+k} \frac{1}{k!} \frac{1}{z+1} \sum_{j=1}^{z+1-k} \sum_{\underline{j}: j_1 + \dots + j_k = z+1-j} x_j \prod_{i=1}^k \frac{x_{j_i}}{j_i} \right) + \frac{(-1)^z}{z+1} x_{z+1}. \end{aligned}$$

Since for every $1 \leq m \leq k+1$

$$\sum_{j=1}^{z+1-k} \sum_{\underline{j}: j_1 + \dots + j_k = z+1-j} x_j \prod_{i=1}^k \frac{x_{j_i}}{j_i} = \sum_{\underline{j}: j_1 + \dots + j_{k+1} = z+1} x_{j_m} \prod_{i \neq m} \frac{x_{j_i}}{j_i} = \sum_{\underline{j}: j_1 + \dots + j_{k+1} = z+1} j_m \prod_{i=1}^{k+1} \frac{x_{j_i}}{j_i},$$

then

$$\begin{aligned} \sum_{j=1}^{z+1-k} \sum_{\underline{j}: j_1 + \dots + j_k = z+1-j} x_j \prod_{i=1}^k \frac{x_{j_i}}{j_i} &= \frac{1}{k+1} \sum_{\underline{j}: j_1 + \dots + j_{k+1} = z+1} \prod_{i=1}^{k+1} \frac{x_{j_i}}{j_i} \sum_{m=1}^{k+1} j_m \\ &= \frac{z+1}{k+1} \sum_{\underline{j}: j_1 + \dots + j_{k+1} = z+1} \prod_{i=1}^{k+1} \frac{x_{j_i}}{j_i}, \end{aligned}$$

and therefore

$$\begin{aligned} p_{z+1} &= \left(\sum_{k=2}^{z+1} (-1)^{z+1+k} \frac{1}{k!} \sum_{\underline{j}: j_1 + \dots + j_k = z+1} \prod_{i=1}^k \frac{x_{j_i}}{j_i} \right) + \frac{(-1)^z}{z+1} x_{z+1} \\ &= \sum_{k=1}^{z+1} (-1)^{z+1+k} \frac{1}{k!} \sum_{\underline{j}: j_1 + \dots + j_k = z+1} \prod_{i=1}^k \frac{x_{j_i}}{j_i}, \end{aligned}$$

which completes the induction and hence proves (8.5).

To show that the definition of x_k from (8.2) coincides with (4.5) we define

$$A_{k,z} := \sum_{\ell=2}^{\infty} \frac{3^{k+z-1}}{(\ell+2)^k(\ell-1)^z},$$

which gives

$$x_k = \frac{1}{3^{2k-1}} A_{k,k}.$$

Thus for $k, z \geq 1, k+z \geq 3$ (otherwise the right side is not defined)

$$A_{k,z} = \sum_{\ell=2}^{\infty} \frac{3^{k+z-2}}{(\ell+2)^{k-1}(\ell-1)^{z-1}} \left(\frac{1}{\ell-1} - \frac{1}{\ell+2} \right) = A_{k-1,z} - A_{k,z-1}.$$

Assume we do not sum to ∞ but to a large finite N such that $A_{1,0}$ and $A_{0,1}$ exist. It can be proved by induction on $k+z$ that

$$A_{k,z} = (-1)^z \left(\sum_{j=1}^{k \vee z} \binom{k+z-j-1}{k-j} A_{j,0} + (-1)^j \binom{k+z-j-1}{z-j} A_{0,j} \right) + \mathcal{O}\left(\frac{1}{N}\right)$$

where $\binom{-1}{0} = 1$. Using this we have

$$A_{k,k} = (-1)^k \left(\sum_{j=1}^k \binom{2k-j-1}{k-j} (A_{j,0} + (1_{\{j \text{ even}\}} - 1_{\{j \text{ odd}\}}) A_{0,j}) \right) + \mathcal{O}\left(\frac{1}{N}\right).$$

So, as

$$A_{0,j} = 3^{j-1} \zeta(j), \quad A_{j,0} = 3^{j-1} (\zeta(j) - b_j), \quad b_j = 1 + \frac{1}{2j} + \frac{1}{3j}$$

we can write, now also for $N = \infty$

$$x_k = \frac{1}{3^{2k-1}} A_{k,k} = \frac{(-1)^k}{3^{2k-1}} \left(\sum_{j=1}^k \binom{2k-j-1}{k-j} 3^{j-1} (-b_j + 1_{\{j \text{ even}\}} 2\zeta(j)) \right)$$

which shows that x_k is of the form (4.5). This completes the proof of the theorem's assertion 3, from which the weights claimed in assertion 4 follow by inspection.

Proof of 1.

To obtain the probability generating function we now calculate

$$\begin{aligned} g(t) &:= \mathbf{E}[t^Z] = \sum_{z=0}^{\infty} t^z \mathbf{P}[Z = z] = \frac{1}{3} \sum_{z=0}^{\infty} \sum_{k=0}^{\infty} \sum_{\underline{j}: j_1 + \dots + j_k = z} \sum (2t)^z \frac{(-1)^{z+k}}{k!} \prod_{i=1}^k \frac{x_{j_i}}{j_i} \\ &= \frac{1}{3} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \sum_{j_1, \dots, j_k} (-2t)^{\sum j_i} \prod_{i=1}^k \frac{x_{j_i}}{j_i} \\ &= \frac{1}{3} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \left(\sum_{j=1}^{\infty} \frac{(-2t)^j x_j}{j} \right)^k = \frac{1}{3} \exp \left(- \sum_{j=1}^{\infty} \frac{(-2t)^j x_j}{j} \right) \end{aligned} \tag{8.6}$$

where we have used (8.5). The sum in the exponential simplifies to

$$\begin{aligned} -\sum_{j=1}^{\infty} \frac{(-2t)^j x_j}{j} &= -\sum_{i=2}^{\infty} \sum_{j=1}^{\infty} \frac{1}{j} \left(\frac{-2t}{(i+2)(i-1)} \right)^j = \sum_{i=2}^{\infty} \log \left(1 + \frac{2t}{(i+2)(i-1)} \right) \\ &= \sum_{i=2}^{\infty} \log \left(\frac{i(i+1) + 2(t-1)}{(i+2)(i-1)} \right) \end{aligned}$$

which proves the formula for the probability generating function.

Proof of 2.

We calculate the first two derivatives of the generating function:

$$\begin{aligned} g'(t) &= g(t) \sum_{i=2}^{\infty} \frac{2}{i(i+1) + 2(t-1)}, \\ g''(t) &= g(t) \left(\left(\sum_{i=2}^{\infty} \frac{2}{i(i+1) + 2(t-1)} \right)^2 - \sum_{i=2}^{\infty} \frac{4}{(i(i+1) + 2(t-1))^2} \right). \end{aligned}$$

So

$$\begin{aligned} \mathbf{E}[Z] &= g'(1) = 1, \\ \mathbf{Var}[Z] &= \mathbf{E}[Z^2] - 1 = \mathbf{E}[Z(Z-1)] = g''(1) = 1 - 4 \sum_{i=2}^{\infty} \frac{1}{i^2(i+1)^2} \end{aligned}$$

and the last assertion follows by

$$\sum_{i=2}^{\infty} \frac{1}{i^2(i-1)^2} = \sum_{i=2}^{\infty} \left(\frac{1}{i} - \frac{1}{i+1} \right)^2 = 2\zeta(2) - (1 + 1 + \frac{1}{4}) - 2\frac{1}{2} = 2\zeta(2) - \frac{13}{4}.$$

Acknowledgements

We thank Richard Hudson, John Wakeley and Steve Evans for pointing out relevant references, and we are grateful to Tom Kurtz for showing us the recent manuscript [DK06] which helped us improve Theorem 2. Part of our work was done at the Erwin Schrödinger Institute in Vienna, whose hospitality is gratefully acknowledged.

References

- [Bur56] P.J. Burke. The output of a queueing system. *Operations Research*, 4 (1956), no. 6, 699-704.
- [DK96] P. Donnelly and T.G. Kurtz. A countable representation of the Fleming Viot measurable diffusion. *Annals of Probability*, 24(2):698-742, 1996.
- [DK99] P. Donnelly and T.G. Kurtz. Particle representations for measure-valued population models. *Annals of Probability*, 27(1):166-205, 1999.
- [DK06] P. Donnelly and T.G. Kurtz. The Eve Process. Manuscript, personal communication.
- [Gri80] R. C. Griffiths. Lines of descent in the diffusion approximation of neutral Fisher-Wright models. *Theor. Pop. Biol.*, 17:37-50, 1980.
- [GT03] R. C. Griffiths and S. Tavaré. The genealogy of a neutral mutation. In *Green, P.J., Hjort, N.L. and Richardson, S. (Eds.), Highly Structured Stochastic Systems*, pages 393-413. Oxford University Press, 2003.

- [JK77] N.L. Johnson and S. Kotz. *Urn Models and their applications*. John Wiley & Sons, 1977.
- [Kim71] M. Kimura. Theoretical foundation of population genetics at the molecular level. *Theo. Pop. Biol.*, 2(2):174–208, 1971.
- [Kin82] J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, 1982.
- [Kur98] T. G. Kurtz. Martingale problems for conditional distributions of Markov processes. *Electronic Journal of Probability*, 3, no. 9, pages 1–29, 1998.
- [Lit75] R.A. Littler. Loss of variability at one locus in a finite population. *Math. Biosci.*, 25:151–163, 1975.
- [RBY04] E.M. Rauch and Y. Bar-Yam. Theory predicts the uneven distribution of genetic diversity within species. *Nature*, 431:449–452.
- [STW84] I. W. Saunders, S. Tavaré, and G. A. Watterson. On the genealogy of nested subsamples from a haploid population. *Adv. Appl. Probab.*, 16:471–491, 1984.
- [Taj90] F. Tajima. Relationship between DNA polymorphism and fixation time. *Genetics*, 125:447–454, 1990.
- [Wak05] J. Wakeley. Coalescent theory. An Introduction. Roberts & Company Publishers, Greenwood Village, 2005, and Scion Publishing Ltd (to appear).
- [Wat82a] G.A. Watterson. Mutant substitutions at linked nucleotide sites. *Adv. Appl. Prob.*, 14:166–205, 1982.
- [Wat82b] G.A. Watterson. Substitution times for mutant nucleotides. *J. Appl. Prob.*, 19A: 59–70, 1982.